36-617: Applied Linear Models
Fall 2021
HW03 – Solutions

1. *Please do Sheather, Ch 3, p. 105, #3,* **Part A**. *Remember that the data is in the "0-textbook" folder in the files area on Canvas for this class.*

   (a) *Develop a simple linear regression model based on least squares that predicts advertising revenue per page from circulation (i.e., feel free to transform either the predictor or the response variable or both variables). Ensure that you provide justification for your choice of model.*

   First, a quick look at the data...
   (You do not have to produce any EDA for your answert to part (a)).

```
> magdata <- read.csv("AdRevenue.csv",header=T)
> str(magdata,width=72,strict.width = "cut")
```

```
'data.frame':        70 obs. of  4 variables:
 $ Magazine                    : chr  "People" "Better Homes and Garden"..
 $ PARENT.COMPANY..SUBSIDIARY: chr  "Time Warner, (Time Inc.)" "Mered"..
 $ AdRevenue                   : num  233 397 286 877 304 ...
 $ Circulation                 : num  3.75 7.64 4.07 32.7 3.21 ...
```

```
> par(mfrow=c(1,3))
> hist(magdata$AdRevenue,main="")
> hist(magdata$Circulation,main="")
> plot(AdRevenue ~ Circulation, data=magdata)
```

   The plot is shown in Figure 1. We can see from the figure that there is severe right skewing in both the Circulation and AdRevenue variables.

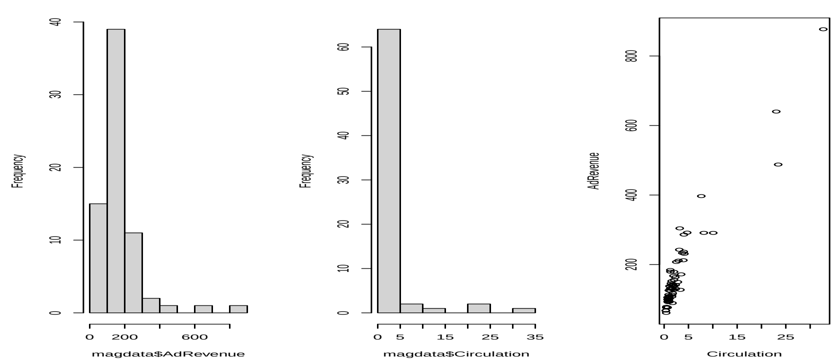


Figure 1: Initial EDA.

   Next, we fit a linear regression model to the untransformed variables:
   (This is a good baseline model to fit when you are considering transformations, but you do not need to include it in your answer to part (a) for this assignment.)

```
> lm.1 <- lm(AdRevenue ~ Circulation, data=magdata)
> summary(lm.1)

Call:
lm(formula = AdRevenue ~ Circulation, data = magdata)

Residuals:
     Min       1Q   Median       3Q      Max
-147.694  -22.939   -7.845   13.810  131.130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99.8095     5.8547   17.05   <2e-16 ***
Circulation  22.8534     0.9518   24.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.22 on 68 degrees of freedom
Multiple R-squared:  0.8945,       Adjusted R-squared:  0.8929
F-statistic: 576.5 on 1 and 68 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(lm.1)
```

The diagnostic plots are shown in Figure 2. We can see that the residuals are also right-skewed and have non-constant variance.

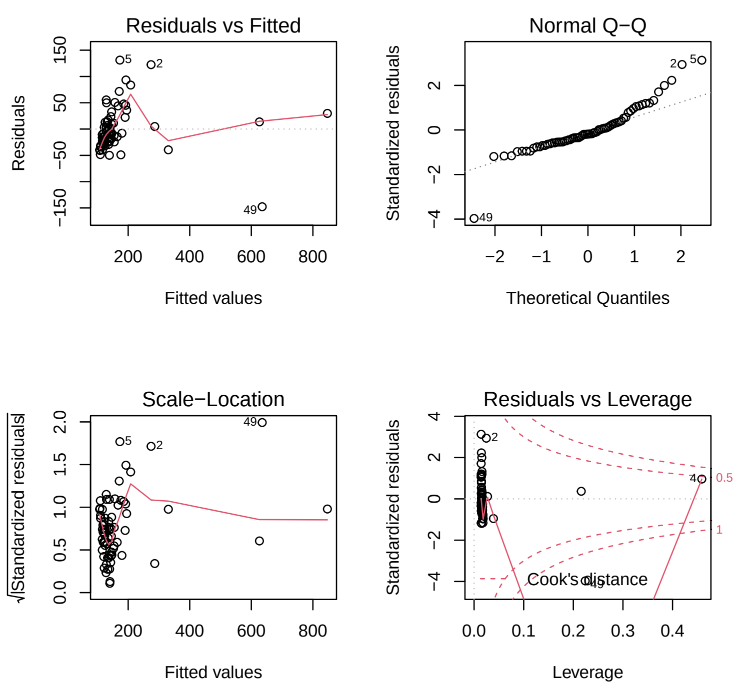


Figure 2: Diagnostics for regression on untransformed variables.

We will try two transformations:
(You should have something like these two analyses in your answer to part (a).)

- Just a log transformation on both variables (since logarithms reduce right-skew, and if it works it gives us an interpretable model);
- The power transformations suggested by Box-Cox.

Here's the "log everything model":

```
> lm.2 <- lm(log(AdRevenue) ~ log(Circulation),data=magdata)
> summary(lm.2)
```

```
Call:
lm(formula = log(AdRevenue) ~ log(Circulation), data = magdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.47022 -0.11142 -0.00532  0.10835  0.42705 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.67473    0.02525  185.16   <2e-16 ***
log(Circulation)  0.52876    0.02356   22.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.1768 on 68 degrees of freedom
Multiple R-squared:  0.881,	Adjusted R-squared:  0.8793 
F-statistic: 503.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(lm.2)
```

The diagnostic plots are shown in Figure 4a, page 5.

For the Box-Cox transformations, I suggest

- First, find the best (rounded) Box-Cox power for x.
- Then, for the model $y \sim$(transformed x), find the best (rounded) Box-Cox power for y.

That way, you are using the information you have learned about x to produce the best possible transformation of y.
First, the suggested transformation for x = Circulation:

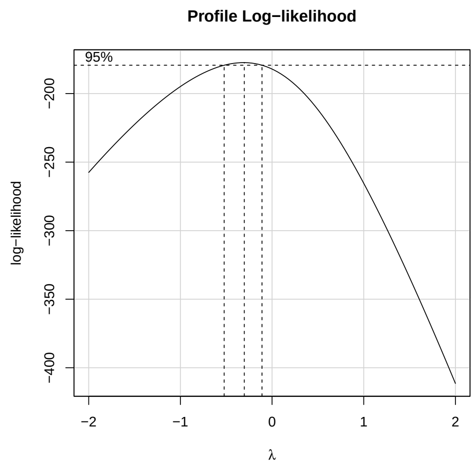
```
> library(car)
> with(magdata,boxCox(Circulation~1))
> with(magdata,powerTransform(Circulation~1)$roundlam)
```

```
  Y1 
-0.5 
```

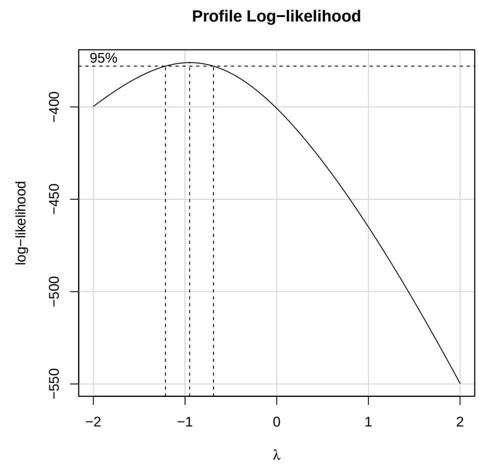
The profile likelihood is shown in Figure 3a, page 4.
Next, the Box-Cox transformation for y = AdRevenue, using the transformed x, $1/\sqrt{\text{Circulation}}$:

```
> lm.3 <- lm(AdRevenue ~ I(Circulation^(-0.5)),data=magdata)
> with(magdata,boxCox(lm.3))
> with(magdata,powerTransform(lm.3)$roundlam)
```

```
Y1 
-1 
```



(a) boxCox profile likelihood for $x \sim 1$. The "rounded" value of λ is -0.5, so the transformation is $1/\sqrt{x}$.



(b) boxCox profile likelihood for $y \sim 1/\sqrt{x}$. The "rounded" value of λ is -1, so the transformation is $1/y$.

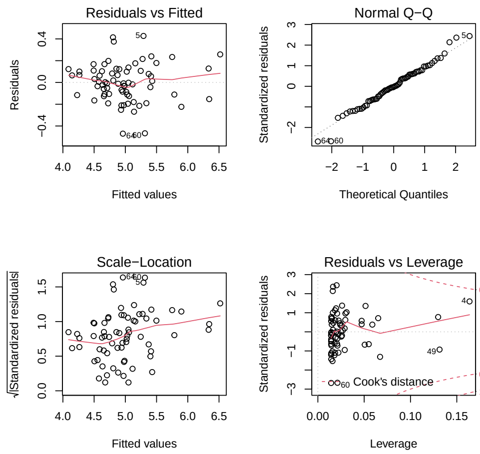
Figure 3: Selecting the boxCox power for $x \sim 1$ and for $y \sim 1/\sqrt{x}$; $x =$ Circulation, $y =$ AdRevenue.

The profile likelihood is shown in Figure 3b.

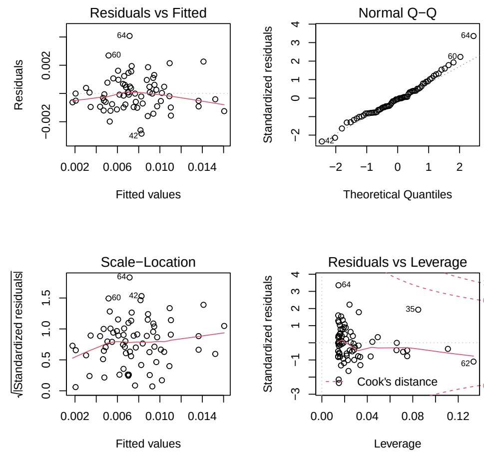
So, the final model suggested by Box-Cox is $1/(\text{AdRevenue}) \sim 1/\sqrt{(\text{Circulation})}$:

```
> magdata$AdRevInv <- 1/magdata$AdRevenue
> magdata$InvSqrtCirc <- 1/sqrt(magdata$Circulation)
> lm.4 <- lm(AdRevInv ~ InvSqrtCirc,data=magdata)
> # the following caused an error in R, which is why I defined the variables above...
> # lm.4 <- lm(I(AdRevenue^(-1)) ~ I(Circulation^(-0.5)),data=magdata)
> summary(lm.4)

Call:
lm(formula = AdRevInv ~ InvSqrtCirc, data = magdata)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0028448 -0.0008745 -0.0000689  0.0006133  0.0040733

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0001662  0.0004000   0.416    0.679
InvSqrtCirc 0.0091424  0.0004571  20.000   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001223 on 68 degrees of freedom
Multiple R-squared:  0.8547,       Adjusted R-squared:  0.8526
F-statistic:   400 on 1 and 68 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(lm.4)
```

The diagnostic plots are shown in Figure 4b.



(a) Regression diagnostics for $\log y \sim \log x$.



(b) Regression diagnostics for $1/y \sim 1/\sqrt{x}$.

Figure 4: Comparing regression diagnostics for the model $\log y \sim \log x$, vs. the "best" boxCox model $1/y \sim 1/\sqrt{x}$; $x =$ Circulation, $y =$ AdRevenue.

(Your choice of model, and justification, should be similar to the following.)
Both models have high R^2's: for lm.2 ($logy \sim logx$), $R^2 = 0.88$ and for lm.4 ($1/y \sim 1/\sqrt{x}$), $R^2 = 0.85$[1]. Comparing coefficient estimates,

```
> summary(lm.2)$coef

                  Estimate Std. Error   t value     Pr(>|t|)
(Intercept)       4.674734 0.02524738 185.15717 1.168946e-93
log(Circulation) 0.528758 0.02356174  22.44138 3.754210e-33
```

```
> summary(lm.4)$coef

                 Estimate   Std. Error    t value     Pr(>|t|)
(Intercept) 0.0001662359 0.0004000311  0.4155575 6.790423e-01
InvSqrtCirc 0.0091424191 0.0004571138 20.0003113 3.417303e-30
```

we see that the slope estimates for both models are highly significantly different from zero, so the regression output doesn't help us distinguish between the models very well. The regression diagnostics for both models are also very similar (Figure 4): in both models the residuals are much more nearly normal, they have nearly constant variance, and there are almost no highly influential data points.

Since both models fit similarly well, I prefer to use the model that is easier to talk about: $\log y \sim \log x$. Referring to the coefficient tables above, we see that for each 1% change in circulation, we can expect about a 0.53% change in ad revenue.

(Here is where you identify which magazines are associated with unusual observations.)
Referring back to Figure 4a, we see that the most extreme residual outliers are observations 5, 60, and 64:

[1]Note that even though the original model lm.1 ($y \sim x$) had an even higher $R^2 = 0.89$, we do not seriously consider it since the regression diagnostic plots (Figure 2) are so bad!

```
> with(magdata,
+      data.frame(Magazine,Circulation,AdRevenue,
+                 StdRes=rstandard(lm.2),leverage=hatvalues(lm.2))[c(5,60,64),])
```

```
             Magazine Circulation AdRevenue    StdRes   leverage
5  Sports Illustrated       3.205   304.185  2.440209 0.02022629
60         Prevention       3.347   127.315 -2.668848 0.02115025
64      Cooking Light       1.717    89.153 -2.678798 0.01432293
```

```
> p <- 1 # number of predictors: x only
> c(leverage.cutoff = 2*(p+1)/dim(magdata)[1])
```

```
leverage.cutoff
     0.05714286
```

We see that *Sports Illustrated* ad revenue overperforms relative to its circulation, while both *Cooking Light* and *Prevention* (a health magazine) underperform. None of these has very high leverage, however (using the rule of thumb that leverage above $4/n$ is "high").

(b) *Find a 95% prediction interval for the advertising revenue per page for magazines with the following circulations:*

i. 0.5 million

ii. 20 million

(Your answer should pretty much go like this.)
This is slightly tricky, because we have to account for the transformation of y in model lm.2 in producing the appropriate prediction interval:

i. 0.5 million

```
> print(pred.i <- predict(lm.2,newdata=data.frame(AdRevenue=0,Circulation=0.5),
+                          interval="prediction"))
```

```
       fit      lwr    upr
1 4.308227 3.947855 4.6686
```

```
> print(interval.i <- exp(pred.i[c(2,3)]))  # Have to un-do the logarithm
```

```
[1]  51.82406 106.54846
```

So, we see that a magazine with circulation of half a million could expect ad revenue between \$51,820 and \$106,550.

ii. 20 million

```
> print(pred.ii <- predict(lm.2,newdata=data.frame(AdRevenue=0,Circulation=20),
+                          interval="prediction"))
```

```
       fit      lwr      upr
1 6.258752 5.885815 6.631689
```

```
> print(interval.ii <- exp(pred.ii[c(2,3)]))  # Have to un-do the logarithm
```

```
[1] 359.8958 758.7626
```

So, a magazine with circulation of 20 million could expect ad revenue between \$359,900 and \$758,760.

(c) *Describe any weaknesses in your model.*

(Your answer should pretty much go like this, though you might also reproduce the summary()'s here.)
We already discussed the summary in part (a): it shows a high R^2 of 0.88, and a highly significant slope estimate (same with the F statistic for overall fit). Referring again to Figure 4a:

- The residuals vs fitted plot doesn't really show any problems. Except for a couple of outliers, that show up more clearly in the QQ plot, there really isn't any vertical pattern or curve that the residuals follow.

- The QQ plot shows that the residuals are following the normal distribution fairly well. There is still a bit of right skewing in the residuals, and two low outliers *Prevention* and *Cooking Light*.
- The scale-location plot doesn't really show any serious problems, though there may be a bit of increasing variance as the predicted ad revenue increases.
- The residuals vs leverage plot shows a couple of high-leverage points, but no really high Cook's Distance values, so these points are not very influential on the fit of the model lm.2; some more detail on the high Cook's Distance points is given in the R output below:

  (You don't have to do the following for your answer, but it is interesting.)

  ```
  > res.lev <- data.frame(magdata$Magazine,StdRes=rstandard(lm.2),
  +                       leverage=hatvalues(lm.2),Cooks.Dist=cooks.distance(lm.2))
  > tail(res.lev[order(cooks.distance(lm.2)),])
       magdata.Magazine     StdRes   leverage Cooks.Dist
  64      Cooking Light -2.6787982 0.01432293 0.05213714
  5  Sports Illustrated  2.4402089 0.02022629 0.06146309
  20    Reader's Digest -1.3103123 0.06716783 0.06181267
  49  AARP The Magazine -0.9279932 0.13138896 0.06513180
  60         Prevention -2.6688479 0.02115025 0.07695149
  4          Parade (1)  1.5938160 0.16374980 0.24870866
  > p <- 1 # number of predictors: x only
  > c(leverage.cutoff = 2*(p+1)/dim(magdata)[1])
  leverage.cutoff
       0.05714286
  ```

Overall, the log-log model seems to fit the data well.

2. *Please do Sheather, Ch 3, p. 105, #3,* **Part B**.

   (a) *Develop a polynomial regression model based on least squares that directly predicts the effect on advertising revenue per page of an increase in circula- tion of 1 million people (i.e., do not transform either the predictor nor the response variable). Ensure that you provide detailed justification for your choice of model. [Hint: Consider polynomial models of order up to 3.]*

   Sheather suggests we try polynomials up to order 3. We will go even higher, to order 5, just to see what happens. To save some space, I will just print out tables of estimated coefficients and standard errors, and R^2's:

   (You might arrive at the cubic model differently.)

   ```
   > lm.5 <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3) +
   +              I(Circulation^4) + I(Circulation^5), data=magdata)
   > ## Note: you could get the same model with
   > ## lm.5 <- lm(AdRevenue ~ poly(Circulation, degree=5, raw=T),data=magdata)
   > summary(lm.5)$r.squared
   [1] 0.9366882
   > summary(lm.5)$coefficients
                         Estimate   Std. Error    t value    Pr(>|t|)
   (Intercept)      57.0382867767 1.737595e+01  3.2825999 0.001668725
   Circulation      47.8288287740 2.213601e+01  2.1606799 0.034469648
   I(Circulation^2)  1.3621060276 7.962225e+00  0.1710710 0.864707640
   I(Circulation^3) -0.6557047334 9.924238e-01 -0.6607104 0.511169347
   I(Circulation^4)  0.0370875190 4.489949e-02  0.8260120 0.411865983
   I(Circulation^5) -0.0005798354 6.533235e-04 -0.8875166 0.378124078
   ```

If we remove some high powers of x, we may find a change in the significance of lower powers (this can happen because powers of x can be *collinear*). So we will also try the model with powers up to order 4 and order 3 only.

```
> lm.6 <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3) +
+              I(Circulation^4), data=magdata)
> summary(lm.6)$r.squared
```

```
[1] 0.935909
```

```
> summary(lm.6)$coefficients
```

```
                    Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)      45.686665431 11.742208057  3.890807 2.377459e-04
Circulation      65.380572047  9.928749452  6.584976 9.325058e-09
I(Circulation^2) -5.499586772  1.900414956 -2.893887 5.173919e-03
I(Circulation^3)  0.220172602  0.104527351  2.106364 3.903839e-02
I(Circulation^4) -0.002733043  0.001694505 -1.612886 1.116144e-01
```

```
> lm.7 <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3),
+            data=magdata)
> summary(lm.7)$r.squared
```

```
[1] 0.933344
```

```
> summary(lm.7)$coefficients
```

```
                   Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)      59.17036829 8.345045881  7.090478 1.118099e-09
Circulation      51.23581639 4.711234296 10.875243 2.334496e-16
I(Circulation^2) -2.50537894 0.411411261 -6.089719 6.476556e-08
I(Circulation^3)  0.05222479 0.009229702  5.658339 3.574381e-07
```

All three models have $R^2 \approx 0.93$, but only model lm.7, with powers just up to order 3, has all of its $\hat{\beta}$'s significiantly different from zero. Since this seems like the best model so far, we consider diagnostic plots for it:
(Always good to look at diagnostic plots!)

```
> par(mfrow=c(2,2))
> plot(lm.7)
```

The plots appear in Figure 5. They suggest that this model isn't doing as well[2] as the log-log model above!

(b) *Find a 95% prediction interval for the advertising page cost for magazines with the following circulations:*

   i. 0.5 million

   ii. 20 million

(Your answer should pretty much go like this.)
This goes just like the predictions we did earlier, except that now since there is no transformation on y, we can use the prediction intervals directly.

[2] We could improve things a bit by fitting $\log y \sim x + x^2 + x^3$ rather than $y \sim x + x^2 + x^2$ but it is still not better than the log-log model we chose above (try it!). Box-Cox on y also does not help, and anyway these ideas go beyond what Sheather is asking for.

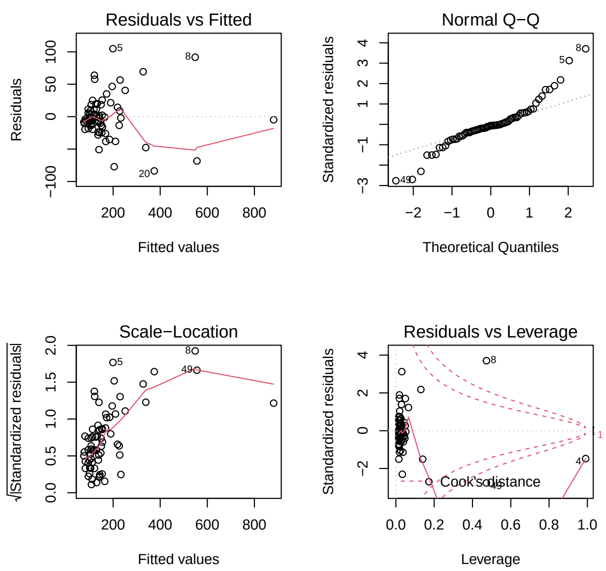


Figure 5: Diagnostics for the 3^{rd} order polynomial model.

```
> print(interval.i <- predict(lm.7,newdata=data.frame(AdRevenue=0,Circulation=0.5),
+                          interval="prediction")[c(2,3)])
```

```
[1]  14.92314 153.41378
```

```
> print(interval.ii <- predict(lm.7,newdata=data.frame(AdRevenue=0,Circulation=20),
+                          interval="prediction")[c(2,3)])
```

```
[1] 418.1790 580.8878
```

So from this model

i. A magazine with a circulation of half a million could anticipate ad revenue between $14,920 and $153,410.
ii. A magazine with a circulation of 20 million could anticipate ad revenue between $418,180 and $580,890.

(c) *Describe any weaknesses in your model.*

(Again, you might copy the model summary here, but otherwise your answer should be pretty similar to the below.)

The summary of model lm.7 above shows an $R^2 = 0.93$, and all significant predictors, which is great. Referring to Figure 5, we see, however, that

- The residuals vs fitted plot shows a lot of right-skew in the fitted values, pretty good symmetry of the residuals around zero, but some pretty large outliers.
- The QQ plot shows a that both the right and left tails or the residual distribution are longer than the normal distribution's tails, with the right tail even longer than the left, and some very large outliers.
- The scale-location plot doesn't show much stability around 1, but part of the problem may be that the fitted values are so skewed-right that there just isn't much data on the right side of the plot from which to make the (red) loess line.
- The residuals vs. leverage plot shows some very large outliers with large leverage values; these also have large Cook's distances, which suggests that they are influential on the fit of lm.7. A few details on large outliers/leverage points are given below.

```
> res.lev <- data.frame(magdata$Magazine,StdRes=rstandard(lm.7),
+                        leverage=hatvalues(lm.7),Cooks.Dist=cooks.distance(lm.7))
> tail(res.lev[order(cooks.distance(lm.7)),],n=7)
```

```
            magdata.Magazine    StdRes   leverage  Cooks.Dist
5        Sports Illustrated  3.125952 0.03102333  0.07821313
46          American Profile -1.509367 0.14002065  0.09273286
2  Better Homes and Gardens  2.179137 0.13037457  0.17797941
20           Reader's Digest -2.702626 0.17242129  0.38044641
49        AARP The Magazine -2.765478 0.47272305  1.71414898
8                USA Weekend  3.705447 0.47304388  3.08140011
4                 Parade (1) -1.478471 0.99099778 60.15734624
```

```
> p <-3 # Number of predictors: x, x^2, X^3
> c(leverage.cutoff = 2*(p+1)/dim(magdata)[1])
```

```
leverage.cutoff
      0.1142857
```

Several magazines have outlying residuals or high leverage; three magazines have high Cook's distances, with higher influence on the fit of lm.7. The last of these, *Parade*, has leverage near one ($h_{ii} = 0.99$!), and a whopping 60 for Cook's Distance!

Overall, this is not a great model. Although the regression output is good, the diagnostic plots reveal many problems with the fit.

3. *Please do Sheather, Ch 3, p. 105, #3,* **Part C**.

   (a) *Compare the model in Part A with that in Part B. Decide which provides a better model. Give reasons to justify your choice.*

   (Your answer should go pretty much like this.
   All of the models have high R^2's and significant predictors (except for some predictors in the higher order polynomial models) . The differences really come in the diagnostic plots. Referring to Figures 4 and 5, as well as our summaries for Parts A(c) and B(c) above, we see that

   - The residual diagnostic plots for the polynomial regression model (Figure 5 show severe model deficiencies, whereas the plots for either the log-log or Box-Cox models Figure 4 show closer agreement with the regression assumptions. So we should pick one of the models from Part A.
   - The two models from Part A perform about equally well, but the log-log model is easier to explain and talk about than the Box-Cox model.

   For these reasons, I prefer the log-log model (lm.2) to all the others tried.

   (b) *Compare the prediction intervals in Part A with those in Part B. In each case, decide which interval you would recommend. Give reasons to justify each choice.*

   (Your answer should be pretty much like this, although if you don't go into an explanation for the different interval widths, that's OK.)
   Here is a table comparing the prediction intervals:

	Circulation = 0.5 million		Circulation = 20 million	
Model	Low Endpoint	High Endpoint	Low Endpoint	High Endpoint
lm.2 (log-log)	51.82	106.55	359.9	758.76
lm.7 (polynomial)	14.92	153.41	418.18	580.89

   For the lower circulation, the log-log interval is narrower. For the higher circulation, the polynomial interval is narrower. We can say a little more: Both models give wider intervals for higher circulations; this is just because higher circulations are farther from the average circulation, and $\mathrm{SE}_{pred}(\hat{y}) = S\sqrt{1 + (x - \overline{x})^2/SXX}$ increases with this distance. The width of the log-log intervals grows faster because we also had to exponentiate the endpoints to get from log($$) intervals back to just $$ intervals. Calculations below:

```
> ## polynomial interval widths:
> data.frame(circ0.5mil=c(width = 153.41 - 14.92), circ20mil = c(width = 580.89 - 418.18))

      circ0.5mil circ20mil
width     138.49    162.71

> ## log-log interval widths:
> data.frame(circ0.5mil = c(width = 106.55 - 51.82), circ20mil = c(width = 758.76 - 359.9))

      circ0.5mil circ20mil
width      54.73    398.86
```

   I prefer the log-log intervals (even though the second one is quite wide) because they are based on a better-fitting model.

4. *Write a brief IDMRAD paper based on your answers to problems 1–3. Remember to label the* **Introduction**, **Data**, **Methods**, **Results** *and* **Discussion** *and* **Technical Appendix** *sections.*

   This report appears at the end of these solutions!

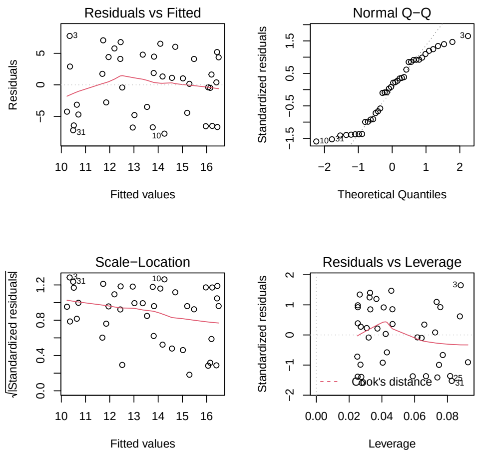
5. *[Based on Gelman & Hill. Ch 3, #1, p. 49] The file* `pyth.dat`*, in the same folder as this hw, contains outcome* `y` *and inputs* `x1`*,* `x2` *for 40 data points, with a further 20 points with the inputs but no observed outcome (for this problem we will ignore these last 20 points). Save the file to your working directory and read it into R using the* `read.table()` *function.*

   (a) *Fit the two models*

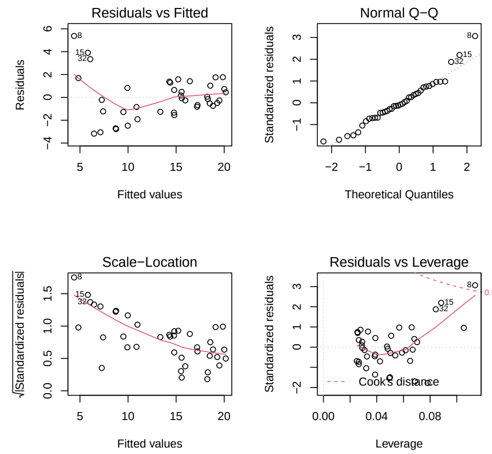
$$\begin{aligned}\mathbf{M1}: \mathtt{y} &= \beta_0 + \beta_1 \mathtt{x1} + \varepsilon \\ \mathbf{M2}: \mathtt{y} &= \beta_0 + \beta_1 \mathtt{x2} + \varepsilon\end{aligned}$$

   Which model provides a better fit for `y`*? Why?*

```
> gh.data <- read.table("pyth.dat",header=T)
> gh.data <- gh.data[apply(gh.data,1,function(x) {!any(is.na(x))}),]
> str(gh.data)

'data.frame':        40 obs. of  3 variables:
 $ y : num  15.68 6.18 18.1 9.07 17.97 ...
 $ x1: num  6.87 4.4 0.43 2.73 3.25 5.3 7.08 9.73 4.51 6.4 ...
 $ x2: num  14.09 4.35 18.09 8.65 17.68 ...

> M1 <- lm(y ~ x1, data=gh.data)
> M2 <- lm(y ~ x2, data=gh.data)
> summary(M1)

Call:
lm(formula = y ~ x1, data = gh.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7409 -4.5056  0.7114  4.3739  7.7547 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.0633     1.5526   6.481 1.25e-07 ***
x1            0.6559     0.2499   2.625   0.0124 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.921 on 38 degrees of freedom
Multiple R-squared:  0.1535,        Adjusted R-squared:  0.1312 
F-statistic:  6.89 on 1 and 38 DF,  p-value: 0.01242

> summary(M2)

Call:
lm(formula = y ~ x2, data = gh.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1751 -1.2352 -0.1867  1.0899  5.3755 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.78532    0.66037   5.732 1.33e-06 ***
x2           0.83223    0.05017  16.589  < 2e-16 ***
```



(a) Diagnostic plots for M1.



(b) Diagnostic plots for M2.

Figure 6: Diagnostic plots for problem 5a.

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.863 on 38 degrees of freedom
Multiple R-squared:  0.8787,       Adjusted R-squared:  0.8755
F-statistic: 275.2 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
> ##
> par(mfrow=c(2,2))
> plot(M1)
> plot(M2)
```

The plots appear in Figure 6. The R^2 is much lower for M1 (0.1535) than for M2 (0.8787). Neither set of residual diagnostic plots looks great: the residual vs fitted and scale-location plots somewhat favor M1, and the QQ plots somewhat favor M2. The Cook's distances are a bit better for M1 also. For R^2 and normality of residuals, I prefer M2.

(b) *Construct new variables* `y2 = y`^{2}, `x12 = x1`^{2}, *and* `x22 = x2`^{2} *and fit the models*

$$\begin{aligned} \mathbf{M3}: \texttt{y2} &= \beta_0 + \beta_1 \texttt{x12} + \varepsilon \\ \mathbf{M4}: \texttt{y2} &= \beta_0 + \beta_1 \texttt{x22} + \varepsilon \end{aligned}$$

Compare the fits of these two models to the models in part (a). Which fits best? Why?

```
> attach(gh.data)
> y2 <- y^2
> x12 <- x1^2
> x22 <- x2^2
> detach()
> M3 <- lm(y2 ~ x12,data=gh.data)
> M4 <- lm(y2 ~ x22,data=gh.data)
> summary(M3)
```

```
Call:
lm(formula = y2 ~ x12, data = gh.data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-189.324 -125.674    4.988  131.052  214.089 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 161.7831    31.7873   5.090    1e-05 ***
x12           1.2971     0.6242   2.078   0.0445 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 131.1 on 38 degrees of freedom
Multiple R-squared:  0.102,	Adjusted R-squared:  0.07841 
F-statistic: 4.318 on 1 and 38 DF,  p-value: 0.04452
```

```
> summary(M4)
```

```
Call:
lm(formula = y2 ~ x22, data = gh.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.280 -31.224  -7.463  25.422  59.571 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  35.1583     9.0306   3.893 0.000387 ***
x22           1.0198     0.0419  24.338  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 33.96 on 38 degrees of freedom
Multiple R-squared:  0.9397,	Adjusted R-squared:  0.9381 
F-statistic: 592.3 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(M3)
> plot(M4)
```

The plots are in Figure 7. Model M4 has the highest R^2 (0.9397), and has residuals vs fitted and scale-location plots that are at least as good as any of the others; on the other hand, we seem to be losing normality of the residuals. Nevertheless I prefer M4 so far.

(c) *To fit the model*

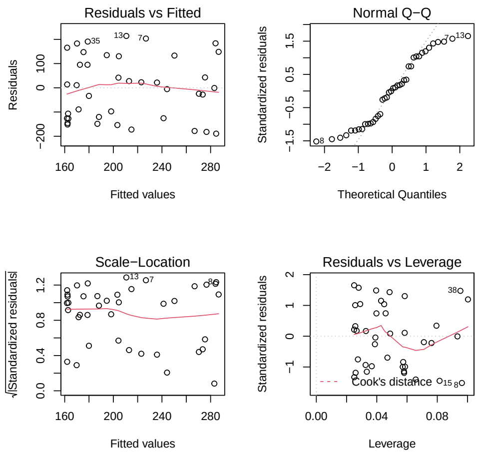
$$\mathtt{y2} = \beta_0 + \beta_1 \mathtt{x1} + \beta_2 \mathtt{x2} + \varepsilon ,$$

we just expand the R modeling language a little bit: `y ~ x1 + x2`*. Fit both of the models*

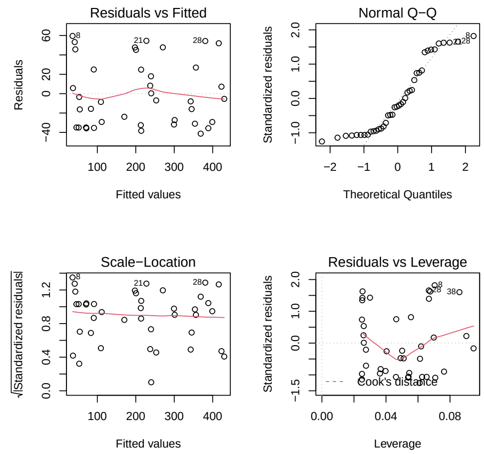
$$\begin{aligned} \mathbf{M5}: \ \mathtt{y} &= \beta_0 + \beta_1 \mathtt{x1} + \beta_2 \mathtt{x2} + \varepsilon \\ \mathbf{M6}: \ \mathtt{y2} &= \beta_0 + \beta_1 \mathtt{x12} + \beta_2 \mathtt{x22} + \varepsilon \end{aligned}$$

Compare these to the earlier models. Which fits best? Why?

```
> M5 <- lm(y ~ x1 + x2, data=gh.data)
> M6 <- lm(y2 ~ x12 + x22, data=gh.data)
> summary(M5)
```



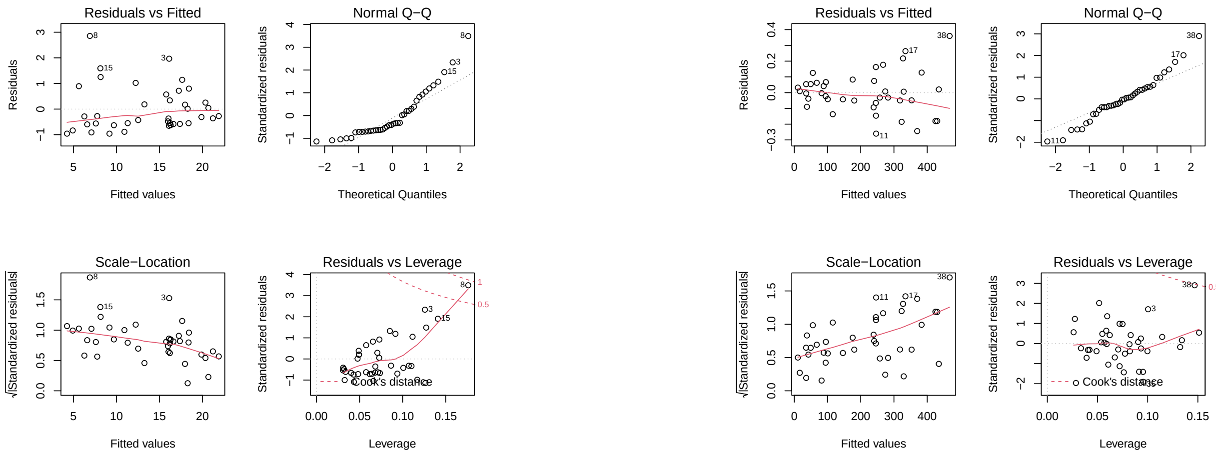
(a) Diagnostic plots for M3.



(b) Diagnostic plots for M4.

Figure 7: Diagnostic plots for problem 5b.

```
Call:
lm(formula = y ~ x1 + x2, data = gh.data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9585 -0.5865 -0.3356  0.3973  2.8548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.31513    0.38769   3.392  0.00166 **
x1           0.51481    0.04590  11.216 1.84e-13 ***
x2           0.80692    0.02434  33.148  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9 on 37 degrees of freedom
Multiple R-squared:  0.9724,        Adjusted R-squared:  0.9709
F-statistic: 652.4 on 2 and 37 DF,  p-value: < 2.2e-16
> summary(M6)
Call:
lm(formula = y2 ~ x12 + x22, data = gh.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26020 -0.05391 -0.00396  0.06367  0.35990

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 0.0026691  0.0422669    0.063     0.95
x12         0.9999672  0.0006419 1557.713   <2e-16 ***
```



(a) Diagnostic plots for M5.

(b) Diagnostic plots for M6.

Figure 8: Diagnostic plots for problem 5c.

```
x22          0.9998685  0.0001663 6011.909   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1344 on 37 degrees of freedom
Multiple R-squared:      1,        Adjusted R-squared:      1
F-statistic: 2.013e+07 on 2 and 37 DF,  p-value: < 2.2e-16
> par(mfrow=c(2,2))
> plot(M5)
> plot(M6)
```

The plots appear in Figure 8.

Putting both x1 and x2 in the model for y really improved the model: M5 has an R^2 of 0.9724, and both predictors are significant (have coefficient estimates significantly different from zero). However the residual diagnostic plots don't look great; in particular it seems like the residuals have a lot of right skew, and leverage seems tio increase with the size of the standardized resduals.

M6 is really winning, though: $R^2 = 1$, and the QQ plot shows good agreement between the residuals and the normal distribution. There seems to be some evidence for non-constant variance though: the residuals vs fitted plot fans out as the fitted valuess increase, and the scale-location plot tells a similar story. On the other hand, only one data point seems to have a concerning Cook's distance.

Based on all of this I like M6 best. Looking at the estimated coefficients for M6, I notice something interesting: $\hat{\beta}_0$ is indistinguishable from 0, and both $\hat{\beta}_1$ and $\hat{\beta}_2$ equal 1, to at least two decimal places (even if we compute the 95% CI's!).

(d) *Can you find a simple, recognizable function* `x3 = (something involving both x1 and x2)`*, so that*

$$\textbf{M7}: \ \mathtt{y} \ = \ \beta_0 + \beta_1 \mathtt{x3} + \varepsilon$$

provides a fit comparable to the best fitting models above? What is going on?

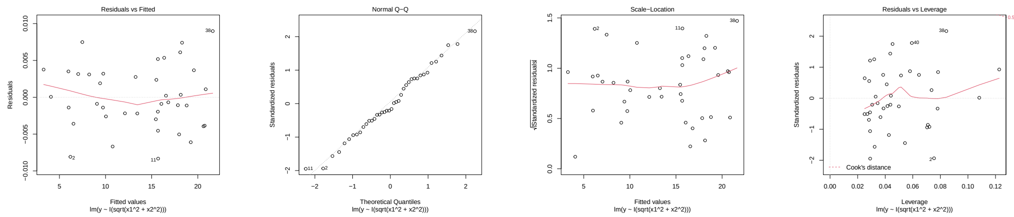


Figure 9: Diagnostic plots for model M7 (problem 5d).

In problem 5c we saw that the model M6 was very nearly

$$y^2 = x1^2 + x2^2 + \varepsilon$$

If we ignore ε, take square roots, and put ε and some "unknown" regression coefficients back in, we get a model like

$$y = \beta_0 + \beta_1 \sqrt{x1^2 + x2^2} + \varepsilon$$

i.e., y is the distance to the origin from some points $(x1, x2)$ in Cartesian space.
Let's try fitting this model:

```
> M7 <- lm(y ~ I(sqrt(x1^2 + x2^2)),data=gh.data)
> summary(M7)
```

```
Call:
lm(formula = y ~ I(sqrt(x1^2 + x2^2)), data = gh.data)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0083283 -0.0027000 -0.0007907  0.0031643  0.0089809

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)          0.0018422  0.0019159    0.962    0.342
I(sqrt(x1^2 + x2^2)) 0.9998313  0.0001316 7596.431   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00434 on 38 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1
F-statistic: 5.771e+07 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(1,4))
> plot(M7)
```

The residual diagnostic plots are in Figure 9. This seems to confirm our suspicions! $R^2 = 1$, the estimated regression coefficients are essentially $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$, and the diagnostic plots looks great:

- The residual vs fitted plot shows little vertical structure.
- The QQ plot shows good adherence to normality.
- The scale-location plot is consistent with constant-variance residuals.
- None of the data points has Cook's distance above 0.5.

Understanding the Relationship Between Circulation Size and Ad Revenue for a Selection of U.S. Consumer Magazines

Brian Junker, Department of Statistics and Data Science

brian@stat.cmu.edu

1 Introduction

The price of advertising (and hence revenue from advertising) is different from one consumer magazine to another. Publishers of consumer magazines argue that magazines that reach more readers create more value for the advertiser. Thus, circulation is an important factor that affects revenue from advertising. In this report, we investigate the relationship between circulation and gross advertising revenue.

In particular we will

- Develop regression models to predict gross advertising revenue per advertising page in 2006 (in thousands of dollars) from circulation (in millions); and
- Illustrate the effect of circulation on ad revenue with two prediction intervals.

2 Data

The data are for the top 70 US magazines ranked in terms of total gross advertising revenue in 2006. The data were obtained from http://adage.com and are given in the file AdRevenue.csv which is available on the book web site.

The variables in the data set are shown in Table 1.

Variable	Definition & Comments
Magazine	The name of each magazine for which data was collected
PARENT.COMPANY	The parent company or subsidiary which publishes this magazine
AdRevenue	The magazine's revenue per advertising page in 2006 (in thousands of dollars)
Circulation	The number of subscribers (in millions) to this magazine

Table 1: Variable Definitions for the `AdRevenue.csv` data set.

Summary statistics for the two quanitative variables are given in Table 2. Further EDA in Appendix A (page 5 below) shows that both variables are substantially skewed right, but that an increasing relationship between the variables is plausible.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Circulation	0.331	0.99225	1.6755	3.118471	2.74325	32.700
AdRevenue	61.101	104.85050	133.7940	171.077200	179.39750	876.907

Table 2: Summary Statistics for AdRevenue and Circulation.

3 Methods

To develop a regression model to predict AdRevenue from Circulation, we considered regression models using logarithmic and Box-Cox power transforms of the variables AdRevenue and Circulation (Appendix B) as well as regression of AdRevenue on polynomial functions of Circulation (Appendix C), up to order 5. We chose our final model based on a summary of each regression analysis and an examiniation of residual diagnostic plots.

With our final model, we calculated AdRevenue intervals in which we would expect to find 95% of companies with circulations of 0.5 million subscribers and 20 million subscribers, respectively, accommodating transformation of variables, if any.

4 Results

We considered regressions using the original variables AdRevenue and Circulation (details in Appendix B, p. 6), logarithmic and power transformations (Appendix B, pp. 7ff.) and polynomial regression using polynomials in Circulation of orders 3, 4 and 5 (Appendix C). All approaches produced models with high R^2 values and highly significant predictors of AdRevenue.

Logarithmic and Power Transformations

Among models with logarithmic and power transformations, the models with the best residual diagnostic plots (Appendix B, pages 7 and 11) were the following two models, shown with estimated regression coefficients:

$$\log(\text{AdRevenue}) = 4.67 + 0.53 \cdot \log(\text{Circulation}) + \varepsilon \tag{1}$$

and

$$1/(\text{AdRevenue}) = 0.0002 + 0.0091 \cdot 1/\sqrt{(\text{Circulation})} + \varepsilon \tag{2}$$

Models (1) and (2) had similar R^2 values (0.881 and 0.8547, respectively) and similarly good residual diagnostic plots. Because the log-log model is more easily interpreted in terms of percent-change, we prefer model (1). Table 3 gives the full table of estimated coefficients and standard errors for model (1).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.674734	0.02524738	185.15717	1.168946e-93
log(Circulation)	0.528758	0.02356174	22.44138	3.754210e-33

Table 3: Estimated coefficients and standard errors for model (1).

Polynomial Regression Models

We fitted polynomial regression models of order 3, 4 and 5, and found that the model of order 3, shown here with estimated regression coefficients,

$$(\text{AdRevenue}) = 59.17 + 51.24 \cdot (\text{Circulation}) - 2.51 \cdot (\text{Circulation})^2 + 0.05 \cdot (\text{Circulation})^3 + \varepsilon \tag{3}$$

was the most successful—all the predictors were significant, and $R^2 \approx 0.93$. Table 4 gives the estimated coefficients and standard errors for this model.

Nevertheless, the residual diagnostic plots for this model (Appendix C, p. 13) did not look as good as the diagnostics for the log-log model (1).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	59.17036829	8.345045881	7.090478	1.118099e-09
Circulation	51.23581639	4.711234296	10.875243	2.334496e-16
I(Circulation^2)	-2.50537894	0.411411261	-6.089719	6.476556e-08
I(Circulation^3)	0.05222479	0.009229702	5.658339	3.574381e-07

Table 4: Estimated coefficients and standard errors for model (3).

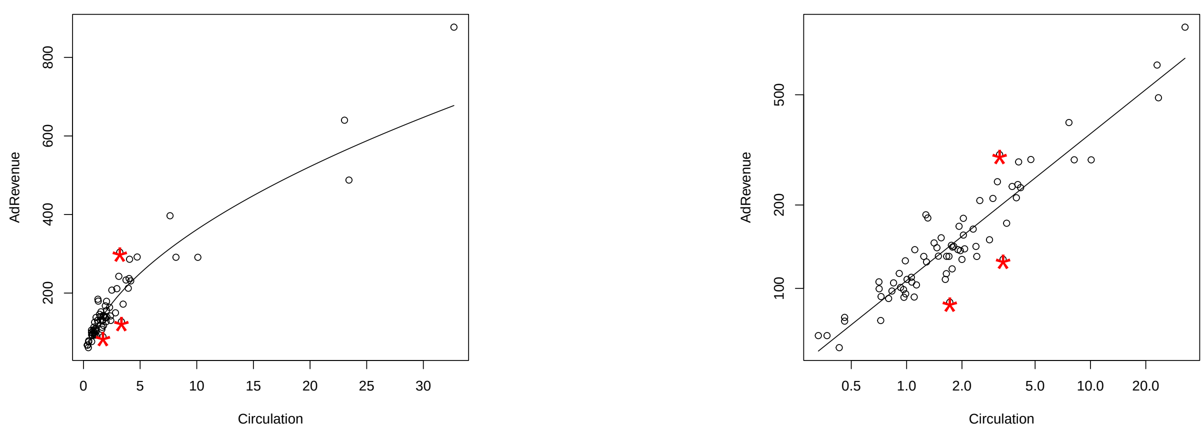
	Magazine	Circulation	Predicted.AdRevenue	Actual.AdRevenue
5	Sports Illustrated	3.205	198.46	304.185
60	Prevention	3.347	203.06	127.315
64	Cooking Light	1.717	142.68	89.153

Table 5: Magazines with unusually high or low ad revenues (in thousands of dollars), given their circulation sizes (in millions of subscriptions), relative to their predicted ad revenue under model (1). These magazines are marked as red points in Figure 1.

Final Model

All three models (1), (2) and (3) have high R^2's and highly significant predictors. Based on residual diagnostic plots (pages 7, 11 and 13 in the Appendix), we can eliminate model (3), which does not follow the assumptions of the linear model as well as the other two. Models (1) and (2) have very similar residual diagnostics, so we are free to choose based on interpretability.

Since the model with logarithms has a simpler interpretation (a 1% change in Circulation is associated with an expected change of 0.53% in AdRevenue), we chose model (1) as our final model. Figure 1 shows the fitted regression line under model (1), laid over the raw data.



(a) Original scale.

(b) Log-log scale.

Figure 1: The fitted model (1) overlaid on the raw data. Points in red correspond to magazines in Table 5.

The three magazines that followed our final model (1) least well are listed in Table 5; nevertheless they did not influence the fit of the model very much. We can see from the table that *Sports Illustrated* outperforms its predicted ad revenue, and both *Cooking Light* and *Prevention* underperform their predicted ad revenues.

Prediction Intervals

Using our final model model (1), we predict 95% of consumer magazines could expect ad revenues in the following intervals, based on their circulations:

- For a magazine with a circulation of 0.5 million subscriptions, the predicted interval for ad revenue is \$51,820 to \$106,550.
- For a magazine with a circulation of 20 million subscriptions, the predicted interval for ad revenue is \$359,900 to \$758,760.

5 Discussion

Among models we considered (power transformations in simple linear regression, as well as polynomial regression), we found that the model that fits the relationship between Circulation and AdRevenue best is a log-log model, shown here with estimated regression coefficients:

$$\log(\text{AdRevenue}) = 4.67 + 0.53 \cdot \log(\text{Circulation}) + \varepsilon \tag{1}$$

The variable $\log(\text{Circulation})$ is a highly significant predictor of $\log(\text{AdRevenue})$; the variation in predicted $\log(\text{AdRevenue})$ accounts for $R^2 \cdot 100\% = 88.1\%$ of the variation in raw $\log(\text{AdRevenue})$. The model can be interpreed as saying that we expect a 0.53% change in Ad Revenue for every 1% change in Circulation. The relationship in model (1) is illustrated in Figure 1 above.

We also calculated intervals predicting a range of Ad Revenues for magazines with circulations of 0.5 million and 20 million subscriptions. As expected, the larger the circulation, the wider the range of possible ad revenues.

These calculations are helpful in determining whether particular magazines are over- or under-performing what we would expect, and we illustrated this with three magazines with the most unusual ad revenues for their circulation sizes, according to model 1; see Table 5 above.

A key limitation of this work is that the data is quite old, from 2006. This limits the generalizability of the results to the present time: the publishing industry has continued to undergo enormous upheavals due to competition from "free" content available on the internet; although we might still expect a log-log relationship to hold up with more current data, we would expect the estimated regression coeffcients (at least) to change.

It also might be useful to have more than 70 magazines, especially to assess whether the relationship holds up for lower-circulation or lower-revenue magazines, and whether the relationship changes from one magazine genre (e.g. sports magazines) to another (e.g. health magazines).

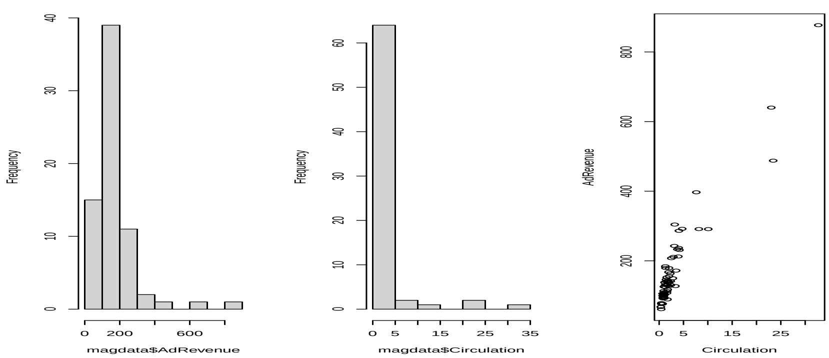
A Initial Look at the Data

We begin by reading in the data and taking a quick look at it:

```
> magdata <- read.csv("AdRevenue.csv",header=T)
> str(magdata,width=72,strict.width = "cut")
```

```
'data.frame':        70 obs. of  4 variables:
 $ Magazine                  : chr  "People" "Better Homes and Garden"..
 $ PARENT.COMPANY..SUBSIDIARY: chr  "Time Warner, (Time Inc.)" "Mered"..
 $ AdRevenue                 : num  233 397 286 877 304 ...
 $ Circulation               : num  3.75 7.64 4.07 32.7 3.21 ...
```

```
> par(mfrow=c(1,3))
> hist(magdata$AdRevenue,main="")
> hist(magdata$Circulation,main="")
> plot(AdRevenue ~ Circulation, data=magdata)
```




```
> rbind(
+ Circulation = summary(magdata$Circulation),
+ AdRevenue = summary(magdata$AdRevenue)
+ )
```

```
              Min.   1st Qu.   Median       Mean   3rd Qu.    Max.
Circulation  0.331   0.99225   1.6755   3.118471   2.74325  32.700
AdRevenue   61.101 104.85050 133.7940 171.077200 179.39750 876.907
```

We see from the plots that both AdRevenue and Circulation are highly skewed-right. However, there does seem to be a linear relationship between these two variables.

B Simple Regression, Transformed Variables

We tried

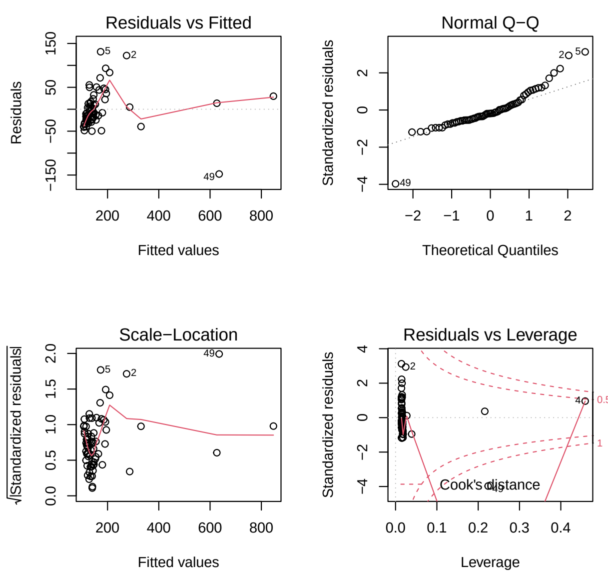
- Simple regression on the original variables: AdRevenue ~ Circulation.
- Simple regression on the logs of the variables: log(AdRevenue) ~ log(Circulation).
- Simple regression with Box-Cox power transformations of the variables; this model turned out to be $1/\text{AdRevenue} \sim 1/\sqrt{\text{Circulation}}$.

The best model turned out to be the log-log model. Some details of our analyses follow:

Original variables

The regression output and residual diagnostic plots for the model AdRevenue ~ Circulation are as follows:

```
> summary(lm.1 <- lm(AdRevenue ~ Circulation,data=magdata))

Call:
lm(formula = AdRevenue ~ Circulation, data = magdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-147.694  -22.939   -7.845   13.810  131.130 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  99.8095     5.8547   17.05   <2e-16 ***
Circulation  22.8534     0.9518   24.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.22 on 68 degrees of freedom
Multiple R-squared:  0.8945,	Adjusted R-squared:  0.8929 
F-statistic: 576.5 on 1 and 68 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(lm.1)
```

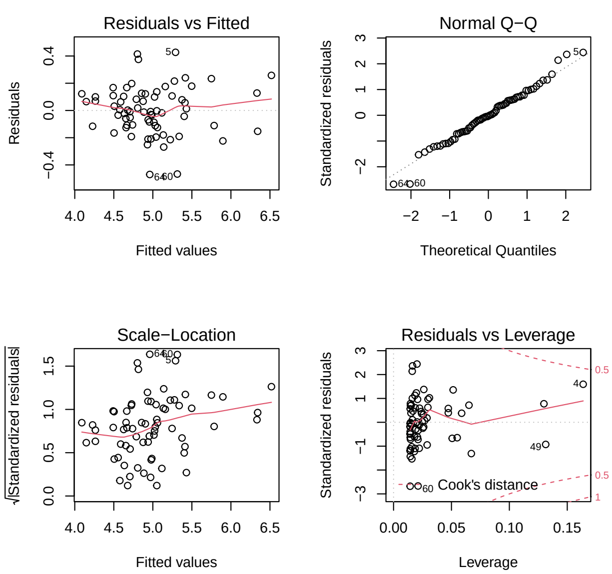


Although $R^2 = 0.8945$ and Circulation is a highly statistically significant predictor, the residual diagnostic plots show skewing in the residuals, to go along with the skewing in AdRevenue and Circulation that we saw in the exploratory plots in Section A.

Log-transformed variables

The regression output and residual diagnostic plots for the model log(AdRevenue) ~ log(Circulation) are as follows:

```
> summary(lm.2 <- lm(log(AdRevenue) ~ log(Circulation),data=magdata))

Call:
lm(formula = log(AdRevenue) ~ log(Circulation), data = magdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.47022 -0.11142 -0.00532  0.10835  0.42705 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.67473    0.02525  185.16   <2e-16 ***
log(Circulation)  0.52876    0.02356   22.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1768 on 68 degrees of freedom
Multiple R-squared:  0.881,	Adjusted R-squared:  0.8793 
F-statistic: 503.6 on 1 and 68 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(lm.2)
```



For this model, $R^2 = 0.881$ is still high, log(Circulation) is still a strong predictor, and the residual diagnostic plots look much better: residuals show no severe vertical patterns, they follow the normal distribution except for a small number of outliers, the location-scale plot shows at most mild violations of non-constant variance, and no data points with high Cook's distances.

We can look at the data points with the highest (but still not concerning) Cook's distances to see what "extreme" data looks like for this model; points with high leverage tend to have small residuals, and vice-versa:

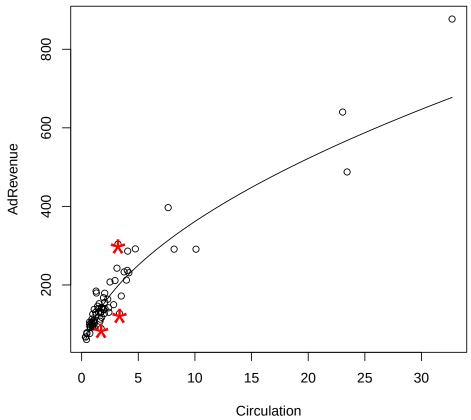
```
> res.lev <- data.frame(Magazine=magdata$Magazine,StdRes=rstandard(lm.2),
+                       leverage=hatvalues(lm.2),Cooks.Dist=cooks.distance(lm.2))
> tail(res.lev[order(cooks.distance(lm.2)),],n=8)

                  Magazine     StdRes   leverage Cooks.Dist
42            Country Home  2.3636899 0.01639929 0.04657547
2  Better Homes and Gardens  1.3575202 0.05146659 0.04999601
64           Cooking Light -2.6787982 0.01432293 0.05213714
5       Sports Illustrated  2.4402089 0.02022629 0.06146309
20          Reader's Digest -1.3103123 0.06716783 0.06181267
49        AARP The Magazine -0.9279932 0.13138896 0.06513180
60              Prevention -2.6688479 0.02115025 0.07695149
4               Parade (1)  1.5938160 0.16374980 0.24870866

> p <- 1 # number of predictors: x only
> c(leverage.cutoff = 2*(p+1)/dim(magdata)[1])

leverage.cutoff
     0.05714286
```

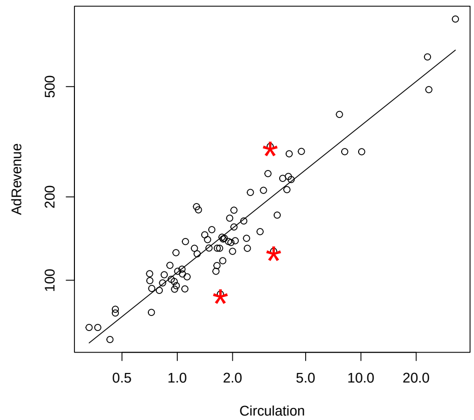
We see that none of these cases have Cook's distances exceeding 0.50. The points that are identified as the three most extreme outliers in the QQ plot correspond to the magazines *Sports Illustrated*, which overperforms expectation, and *Cooking Light* and *Prevention*, which both underperform:

```
> data.frame(Magazine=magdata$Magazine,Circulation=magdata$Circulation,
+            Predicted.AdRevenue=round(exp(predict(lm.2)),2),
+            Actual.AdRevenue=magdata$AdRevenue)[c(5,60,64),]

             Magazine Circulation Predicted.AdRevenue Actual.AdRevenue
5  Sports Illustrated       3.205              198.46          304.185
60         Prevention       3.347              203.06          127.315
64      Cooking Light       1.717              142.68           89.153
```

Finally, here are plots of the fitted regression line, overlaid on the raw data, in the original scale and in a log-log scale. The points colored in red correspond to the "outlier" magazines in the table above.

```
> plot(AdRevenue ~ Circulation,data=magdata)
> regline <- function(x) {exp(4.67 + 0.53*log(x))}
> curve(regline,add=T)
> points(c(3.205,3.347,1.717),c(304.185,127.315,89.153),col="red",pch="*",cex=3)

> plot(AdRevenue ~ Circulation,data=magdata,log="xy")
> regline <- function(x) {exp(4.67 + 0.53*log(x))}
> curve(regline,add=T)
> points(c(3.205,3.347,1.717),c(304.185,127.315,89.153),col="red",pch="*",cex=3)
```



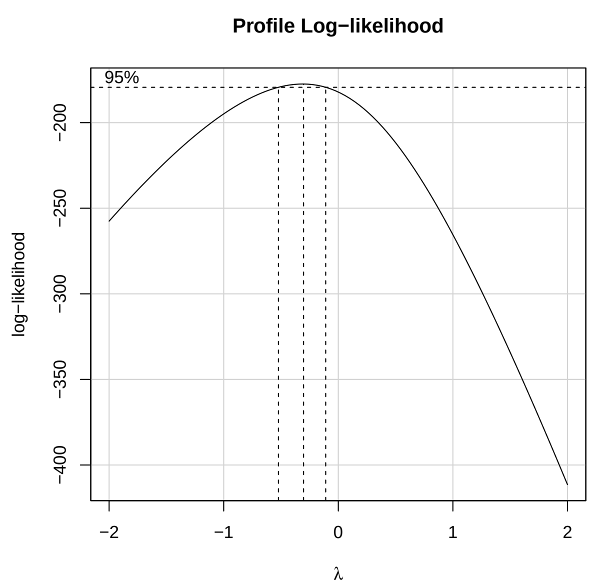


Box-Cox transformed variables

In order to find the Box-Cox transformations, we first find the best transform for x, and then using the transformed x, we find the best transform for y.

First, the suggested transformation for x = Circulation:

```
> library(car)
> with(magdata,powerTransform(Circulation~1)$roundlam)

  Y1
-0.5

> with(magdata,boxCox(Circulation~1))
```

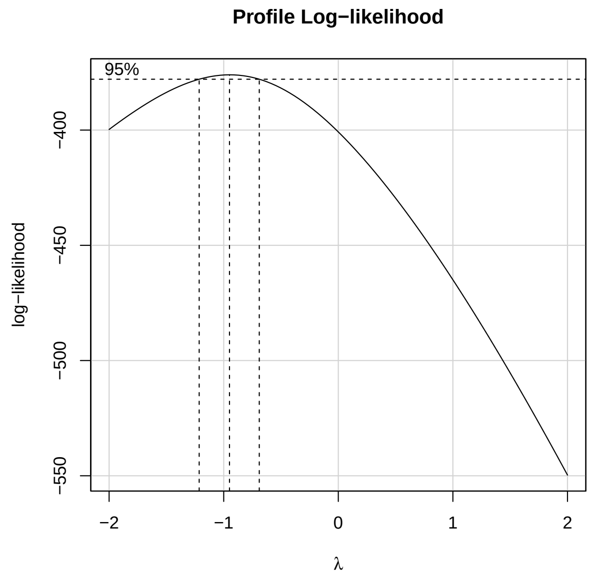


Then, the suggested transform for y = AdRevenue, when regressing on $x = 1/\sqrt{\text{Circulation}}$:

```
> lm.3 <- lm(AdRevenue ~ I(Circulation^(-0.5)),data=magdata)
> with(magdata,powerTransform(lm.3)$roundlam)
```

```
Y1
-1
```

```
> with(magdata,boxCox(lm.3))
```

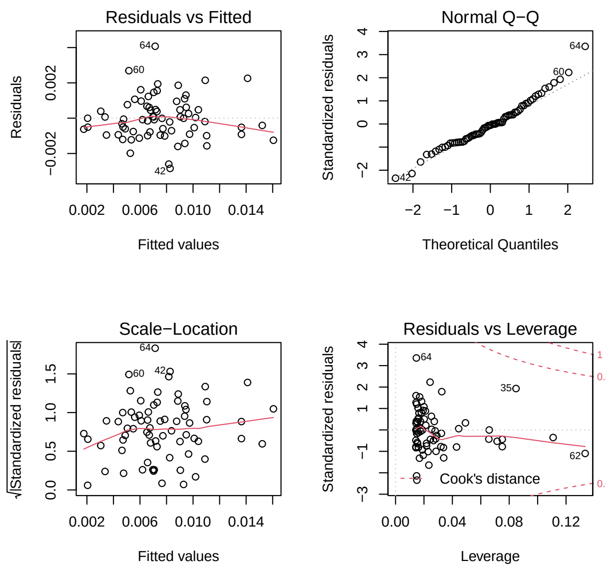


So our final Box-Cox model should be $1/\text{AdRevenue} \sim 1/\sqrt{\text{Circulation}}$:

```
> magdata$AdRevInv <- 1/magdata$AdRevenue
> magdata$InvSqrtCirc <- 1/sqrt(magdata$Circulation)
> lm.4 <- lm(AdRevInv ~ InvSqrtCirc,data=magdata)
> # the following caused an error in R, which is why I defined the variables above...
> # lm.4 <- lm(I(AdRevenue^(-1)) ~ I(Circulation^(-0.5)),data=magdata)
> summary(lm.4)
```

```
Call:
lm(formula = AdRevInv ~ InvSqrtCirc, data = magdata)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0028448 -0.0008745 -0.0000689  0.0006133  0.0040733 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0001662  0.0004000   0.416    0.679    
InvSqrtCirc 0.0091424  0.0004571  20.000   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001223 on 68 degrees of freedom
Multiple R-squared:  0.8547,	Adjusted R-squared:  0.8526 
F-statistic:   400 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(lm.4)
```



This model has an $R^2 = 0.8547$, nearly as good as the log-log model, and again (transformed) Circulation is a strong predictor of (transformed) AdRevenue. The residual diagnostic plots are also very comparable to the corresponding plots for the log-log model.

Conclusions, Simple Regression

Here's a brief comparison of the models (x=Circulation, y=AdRevenue):

Strategy	Model	R^2	Significant Predictor?	Comments on Residual Diagnostics
No Transform	$y \sim x$	0.8945	yes	x, y and residuals all skewed right; some severe outliers and larger Cook's distances.
log-log	$\log(y) \sim \log(x)$	0.881	yes	Assumptions of normality and constant variance for residuals approximately satisfied; few outliers; no large Cook's distances.
Box-Cox	$1/y \sim 1/\sqrt{x}$	0.8547	yes	Similar to log-log diagnostics.

Since there isn't much difference in terms of fit and residual diagnostics between the log-log and Box-Cox models, we should choose based on interpretability. The log-log model has a simpler interpreation: a 1% change in Circulation can be expected to produce a $\hat{\beta}_1 = 0.53\%$ change in AdRevenue. Therefore we prefer the log-log model.

C Polynomial Regression, Untransformed Variables

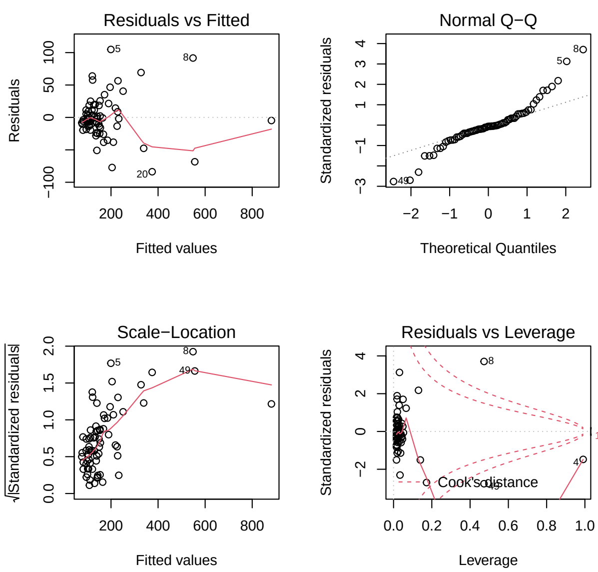
We tried polynomial models of order 5, 4, and 3. To save space, we just quote the R^2 and coefficient tables for each model:

```
> lm.5 <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3) +
+              I(Circulation^4) + I(Circulation^5), data=magdata)
> ## Note: you could get the same model with
> ## lm.5 <- lm(AdRevenue ~ poly(Circulation, degree=5, raw=T),data=magdata)
> summary(lm.5)$r.squared

[1] 0.9366882

> summary(lm.5)$coefficients

                      Estimate   Std. Error    t value    Pr(>|t|)
(Intercept)      57.0382867767 1.737595e+01  3.2825999 0.001668725
Circulation      47.8288287740 2.213601e+01  2.1606799 0.034469648
I(Circulation^2)  1.3621060276 7.962225e+00  0.1710710 0.864707640
I(Circulation^3) -0.6557047334 9.924238e-01 -0.6607104 0.511169347
I(Circulation^4)  0.0370875190 4.489949e-02  0.8260120 0.411865983
I(Circulation^5) -0.0005798354 6.533235e-04 -0.8875166 0.378124078

> lm.6 <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3) +
+              I(Circulation^4), data=magdata)
> summary(lm.6)$r.squared

[1] 0.935909

> summary(lm.6)$coefficients

                     Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)      45.686665431 11.742208057  3.890807 2.377459e-04
Circulation      65.380572047  9.928749452  6.584976 9.325058e-09
I(Circulation^2) -5.499586772  1.900414956 -2.893887 5.173919e-03
I(Circulation^3)  0.220172602  0.104527351  2.106364 3.903839e-02
I(Circulation^4) -0.002733043  0.001694505 -1.612886 1.116144e-01

> lm.7 <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3),
+            data=magdata)
> summary(lm.7)$r.squared

[1] 0.933344

> summary(lm.7)$coefficients

                    Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)      59.17036829 8.345045881  7.090478 1.118099e-09
Circulation      51.23581639 4.711234296 10.875243 2.334496e-16
I(Circulation^2) -2.50537894 0.411411261 -6.089719 6.476556e-08
I(Circulation^3)  0.05222479 0.009229702  5.658339 3.574381e-07
```

All three models have $R^2 \approx 0.93$; in the models of order 4 and 5, the predictors $(\text{Circulation})^4$ and $(\text{Circulation})^5$ were not significant predictors; all predictors were significant for the model of order 3. Therefore we concentrate on the model of order 3. Here is the complete summary and residual diagnostics for that model:

```
> summary(lm.7)
```

```
Call:
lm(formula = AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3),
    data = magdata)

Residuals:
   Min      1Q Median      3Q     Max
-83.75 -13.56  -2.16  11.46 104.82

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      59.17037    8.34505   7.090 1.12e-09 ***
Circulation      51.23582    4.71123  10.875 2.33e-16 ***
I(Circulation^2) -2.50538    0.41141  -6.090 6.48e-08 ***
I(Circulation^3)  0.05223    0.00923   5.658 3.57e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.06 on 66 degrees of freedom
Multiple R-squared:  0.9333,	Adjusted R-squared:  0.9303
F-statistic: 308.1 on 3 and 66 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(lm.7)
```



This set of residual diagnostic plots does not look as good as either the log-log or Box-Cox residual diagnostics in Section B. Although R^2 is higher, we seem to be farther from the assumptions underlying regression here.

D Final Model and Predictions

Comparing the regression output and residual diagnostic plots for the log-log model and polynomial model of order 3:

- The log-log model has dramatically better residual diagnostic plots;
- The log-log model has a significant predictor and an R^2 nearly as high as the polynomial model;
- The log-log model has a simple interpretation: for every 1% increase in Circulation, we can expect a 0.53% increase in Ad Revenue.

For these reasons, we prefer the log-log model.

Here are 95% prediction intervals for Ad Revenue, for a publication with circulation of 0.5 million and 20 million, respectively, from the log-log model. Note that we have to exponentiate the endpoints of the intervals, to “undo” the log transformation on AdRevenue.

```
> int.A.i <- round(exp(predict(lm.2,newdata=data.frame(AdRevenue=0,Circulation=0.5),
+                          interval="prediction")[c(2,3)]),2)
> int.A.ii <- round(exp(predict(lm.2,newdata=data.frame(AdRevenue=0,Circulation=20),
+                          interval="prediction")[c(2,3)]),2)
> data.frame("Lower Endpoint"=c("Circulation 0.5 Million"=int.A.i[1],
+                                "Circulation 20 Million"=int.A.ii[1]),
+            "Upper Endpoint"=c("Circulation 0.5 Million"=int.A.i[2],
+                                "Circulation 20 Million"=int.A.ii[2]))

                        Lower.Endpoint Upper.Endpoint
Circulation 0.5 Million          51.82         106.55
Circulation 20 Million          359.90         758.76
```